

Learning Text-to-Motion from Partially Observed Web Motions via Part-Aware Masked Modeling

Boyuan Li
Renmin University of China
Beijing, China
liboyuan@ruc.edu.cn

Sipeng Zheng
BeingBeyond
Beijing, China
spzheng@beingbeyond.com

Bin Cao
Institute of Automation, Chinese
Academy of Sciences
Beijing, China
caobin2023@ia.ac.cn

Ruihua Song*
Renmin University of China
Beijing, China
rsong@ruc.edu.cn

Zongqing Lu
Peking University
Beijing, China
zongqing.lu@pku.edu.cn

Abstract

Internet videos are a promising complement to motion-capture data for text-to-motion generation, but most web shots do not contain a fully visible body. Existing pipelines either keep all recovered joints as if they were equally reliable targets or discard incomplete clips altogether, yet both clip-level policies are mismatched to web motion with mixed-quality body parts. We study this setting as part-level credible supervision, where observed parts provide reliable learning targets while unobserved parts should not directly define the loss. We propose RoPAR, a part-aware text-to-motion framework that uses only the credible body parts as direct supervision and infers the remaining parts from text and visible context. Specifically, we divide the body into five kinematic parts, estimate part credibility from web videos, learn a latent motion space from the credible part motions, and use masked generation to reconstruct the full motion sequence. We also build K700-M, a dataset of 198k motion-text pairs mined from Kinetics-700 for studying this setting. On K700-M, RoPAR outperforms both clip-level alternatives under matched *Keep-All* and *Discard* protocols, reducing the best baseline FID by 19.1% and improving the best baseline R@1 by 20.3% in the scale-preserving *Keep-All* setting. These results suggest that, for web-recovered motion, keeping partially useful clips and restricting supervision to image-supported body parts is more effective than clip-level filtering.

CCS Concepts

• **Computing methodologies** → **Motion processing**; *Neural networks*.

Keywords

text-to-motion generation, partially observed web motions, part-aware masked modeling

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License. *MM '26, Rio de Janeiro, Brazil*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
<https://doi.org/10.1145/3767308.3835813>

ACM Reference Format:

Boyuan Li, Sipeng Zheng, Bin Cao, Ruihua Song, and Zongqing Lu. 2026. Learning Text-to-Motion from Partially Observed Web Motions via Part-Aware Masked Modeling. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26)*, November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3767308.3835813>

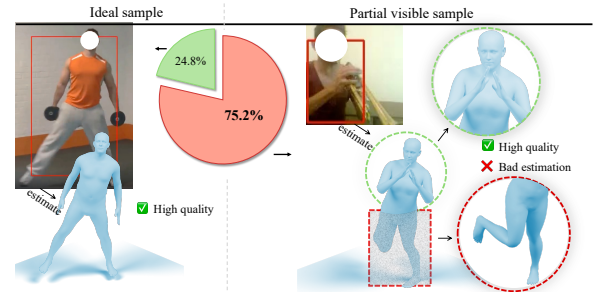


Figure 1: Web-recovered motion is usually only partially observed. Our pilot analysis on Kinetics-700 shows that only 24% of frames contain a fully visible body; in the remaining frames, at least one body part must be hallucinated from incomplete visual evidence.

1 Introduction

Text-to-motion (T2M) models need large motion-text corpora, yet high-quality motion capture data is still limited in scale, scene diversity, and action coverage [13, 24]. Web video can help: it contains open-vocabulary actions, object interactions, and camera view-points that curated MoCap datasets rarely capture. Recent efforts such as Motion-X and Being-M0 scale T2M with web-recovered motion [23, 30, 34, 39]. Our focus is narrower. Once 3D motion is recovered from raw videos with Human Mesh Recovery (HMR) pipelines, the resulting supervision is often only partially credible because not every body part is directly supported by image evidence.

Web videos seldom show the full body for an entire clip. People are cropped, self-occluded, or blocked by objects and other people. HMR still outputs a complete motion sequence in these cases by filling in missing joints with learned priors. The result often looks

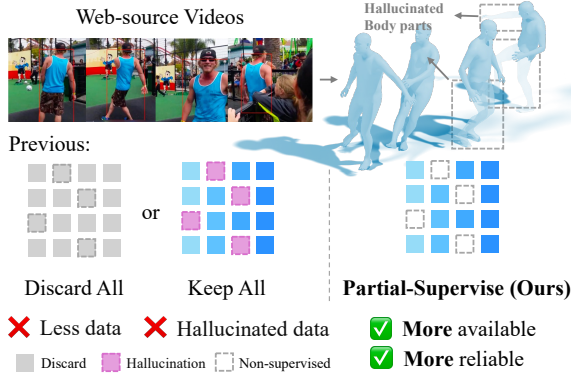


Figure 2: Compared with the two standard clip-level policies for web-recovered motion, RoPAR keeps the clips but supervises only the visible parts. This avoids both wasting useful partial evidence under *Discard* and learning directly from unsupported HMR completions under *Keep-All*.

plausible, but plausibility is not the same as observability: some joints are supported by image evidence, while others are guesses. Treating them as equally reliable targets creates a mismatch between the source video and the supervision signal.

Our pilot analysis on Kinetics-700 [6], summarized in Fig. 1, shows that this issue is common. Only 24% of frames contain a fully visible body. In the remaining 76%, at least one part has to be inferred from incomplete evidence. At the sequence level, more than 70% of K700-M clips contain at least one non-visible body part. Web-recovered motion is therefore usually only partially observed.

One obvious fix is to drop incomplete clips. The problem is that this would remove most web data, including many clips that still contain useful local motion. A talking-head video may still provide reliable upper-body dynamics, and a cooking clip may still reveal informative arm motion. Many web clips remain useful at the part level even when the whole body is not. The challenge is to learn from the observed parts without forcing the model to imitate unsupported ones.

Current T2M pipelines, illustrated in Fig. 2, mostly fall back on two clip-level policies for web-recovered motion. *Keep-All* treats the full recovered body as supervision, even when some parts were never observed. *Discard* removes any clip with a non-visible body part, which means throwing away most of the source pool. Both policies fit web motion poorly because reliability varies within clips and across the dataset. The supervision rule has to work at the part level.

RoPAR is built around a simple learning principle: web-mined T2M should selectively trust the body parts that are supported by image evidence. Reconstructed joints should not all be treated as ground truth. We decompose the body into five kinematic parts, estimate which parts are credible in each frame from 2D pose confidence, and use only those parts as direct supervision. Non-credible parts are masked out of the target but still remain as missing context. A part-aware latent model learns from the credible motions, and a masked generator infers the missing parts from text and the

remaining visible context before a diffusion head refines the full-body sequence. The result is a model that turns partial visibility into a supervision signal and reduces label noise.

To study this setting at scale, we build K700-M from Kinetics-700 [6]. K700-M contains 198,627 motion-text pairs recovered from real web videos and is intentionally harder than MoCap-based datasets because truncation, occlusion, and camera bias are common. We use it as a dataset and evaluation testbed for T2M under partial observability. It is not meant to be a universal ranking benchmark. On K700-M, RoPAR outperforms strong baselines under both standard web-data policies: keeping the full recovered source pool or discarding partially observed clips. We also evaluate on HumanML3D [13] as a fully credible control to check whether the same architecture still works when partial credibility is not an issue. Put differently, RoPAR is aimed at the supervision problem inside web-scale T2M. Other work expands coverage and capacity. We focus on what the model should trust once web-recovered motion is only partly observed.

Our contributions are threefold:

- We formulate web-mined T2M under partial observability as a problem of part-level credible supervision, and show why the two standard clip-level web-data policies, *Keep-All* and *Discard*, are mismatched to this regime.
- We propose RoPAR, a part-aware masked modeling framework that uses only credible body parts as direct supervision while inferring unsupported parts from text and visible context.
- We build K700-M together with a matched evaluation protocol under *Keep-All* and *Discard*, and show that, within this partial-observability setting, selective part-level supervision consistently outperforms both clip-level alternatives and scales better with more web data.

2 Related work

2.1 Text-to-motion generation

Early text-to-motion (T2M) work established the basic problem on curated motion datasets through sequence-to-sequence, GAN, and VAE formulations, including Text2Action [1], Language2Pose [2], Action2Motion [15], ACTOR [26], TEMOS [27], TM2T [14], and MotionCLIP [32]. These methods showed that language can drive plausible 3D motion synthesis, but they were developed under a relatively clean supervision regime in which each training clip provides a reliable full-body target.

Recent progress has mostly come from stronger generative backbones. Diffusion-based approaches such as MotionDiffuse [41], MDM [33], MLD [7], ReMoDiffuse [42], and Make-An-Animation [4] improve motion quality and text alignment by modeling motion as a continuous denoising process. Discrete-token and language-style methods instead quantize motion and use autoregressive or masked prediction, as in T2M-GPT [40], MotionGPT [17], and MoMask [12]. More recent studies also explore part coordination, fine-grained control, and streaming generation, including ParCo [45], FineMoGen [43], MotionStreamer [36], the recent rethinking study on masked autoregression and diffusion [25], and PlanMoGPT [18]. ParCo and FineMoGen are the closest to our setting because explicit part streams make partial supervision expressible: without a

part-wise representation, one cannot trust the visible arms while masking out unsupported legs in the same clip. Prior methods still assume those part targets are reliable once the sequence is given and use part structure mainly for coordination or control. RoPAR uses the same decomposition for a different purpose: deciding which targets are allowed to contribute to the loss.

2.2 Learning from web-scale videos

To overcome the scale limits of MoCap corpora such as AMASS [24] and HumanML3D [13], recent work increasingly turns to internet video as a data source. Motion-X [23] provides a large expressive motion dataset, while more recent vision-language-motion pipelines such as Being-M0 [34], Being-M0.5 [5], Go to Zero [9], and HY-Motion [35] push toward million-scale motion learning and more controllable generation. These works broaden action coverage and visual variety beyond classical motion-capture datasets. Most of them focus on data scale, model scale, or controllability. They spend much less attention on deciding what supervision to trust under partial observability.

Their success also depends on advances in monocular 3D human reconstruction, and these pipelines typically use the recovered full body as the training target. Earlier systems such as SPIN [21], VIBE [19], and PARE [20] improved mesh recovery through optimization in the loop, temporal modeling, and part attention. More recent in-the-wild methods such as SLAHMR [38], Humans in 4D [11], WHAM [30], WHAC [39], and GVHMR [29] better recover camera motion, global trajectory, and world grounding from unconstrained video, while RoHM [44] improves robustness through diffusion-based reconstruction. We do not revisit HMR accuracy or million-scale model design here. We focus on a narrower question inside those pipelines: how should a T2M model learn when different body parts have different levels of visual support?

2.3 Part-aware modeling and imperfect supervision

Part structure has already proved useful for motion generation and control. ParCo [45] coordinates multiple body parts during decoding, FineMoGen [43] studies fine-grained spatio-temporal motion control, and TEACH [3] together with earlier compositional animation work [10] explores how complex instructions can be assembled from simpler motion components. These studies show that part-aware representations help semantic grounding, controllability, and local motion quality. They are also the closest precursors to our setting, because only a part-decomposed model can naturally apply supervision to one body region while withholding it from another. The difference is that earlier part-aware T2M work still assumes each part target is trustworthy once the sequence is given, whereas RoPAR uses visibility to decide which parts are supervised directly and which parts must be inferred.

Our setting is additionally related to learning with imperfect supervision. Robust learning under noisy labels has been studied through peer teaching, sample filtering, and semi-supervised correction, with representative methods such as Co-teaching [16], DivideMix [22], and noisy-label surveys [8]. However, these methods usually treat reliability at the example level: a sample is trusted, reweighted, or discarded as a whole. Web-mined motion is more

structured. Within the same clip, one arm may be well observed while the legs are weakly supported, or vice versa. RoPAR therefore combines part-aware modeling with a reliability-aware training rule: non-visible parts are masked out of direct supervision, and the model learns completion from text together with the remaining visible evidence. This is the key difference between our setting and prior part-aware T2M or generic noisy-label learning.

3 Method

Given a text prompt T and a web-mined motion sequence $\mathcal{M} = \{m_i\}_{i=1}^N$ recovered by an HMR system, our goal is to train a full-body generator even though some body parts in $\mathcal{M}$ are not actually observed in the source video. RoPAR addresses this in three steps: (1) estimate a part-level observability mask from 2D pose confidence, (2) learn a part-aware latent space from only credible motion segments, and (3) train a visibility-guided masked generator that predicts missing part tokens from text and visible context.

The key issue is that masked tokens have two different meanings in our setting. Some credible tokens are deliberately masked so that the model learns to predict them. Other tokens correspond to body parts whose HMR labels are weakly supported by the source video; these tokens are hidden from the input because they are unreliable, but they never serve as prediction targets. RoPAR therefore separates conditioning from supervision: the generator learns from text and the remaining credible evidence, while unreliable parts are excluded from direct supervision.

3.1 Visibility-aware part decomposition

This step converts image evidence into a part-level credibility signal for the recovered 3D motion. We partition the human skeleton into five kinematic parts $P = \{\text{Torso}, \text{L-Arm}, \text{R-Arm}, \text{L-Leg}, \text{R-Leg}\}$. The 3D motion used for training still comes from the HMR reconstruction $\mathcal{M}$, while ViTPose [37] is run on the original RGB frames only to assess whether each body part is sufficiently supported by 2D evidence. Specifically, ViTPose provides a 2D confidence score $C_{i,j} \in [0, 1]$ for each joint j at frame i , and we define the observability mask of part p at frame i as

$$O_i^p = \begin{cases} 1 & \text{if } \frac{1}{|J_p|} \sum_{j \in J_p} C_{i,j} > \tau, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where J_p is the joint set of part p and τ is the visibility threshold.

If $O_i^p = 1$, the part is treated as *credible*. If $O_i^p = 0$, the HMR output for that part is kept in the sequence for bookkeeping but is not trusted as supervision. This distinction lets us keep partially useful clips. It avoids discarding entire sequences.

3.2 Part-aware latent modeling

We encode each part into a continuous latent token with a part-aware variational autoencoder (P-VAE). For part p , the input feature m_i^p contains local joint positions, rotations, and velocities. We also append the root linear velocities (r^x, r^z) and angular velocity r^a so that each part remains grounded in the global body trajectory. The encoder produces a latent token $z_i^p = E_\phi(m_i^p)$, and the decoder reconstructs $\hat{m}_i^p = D_\psi(z_i^p)$.

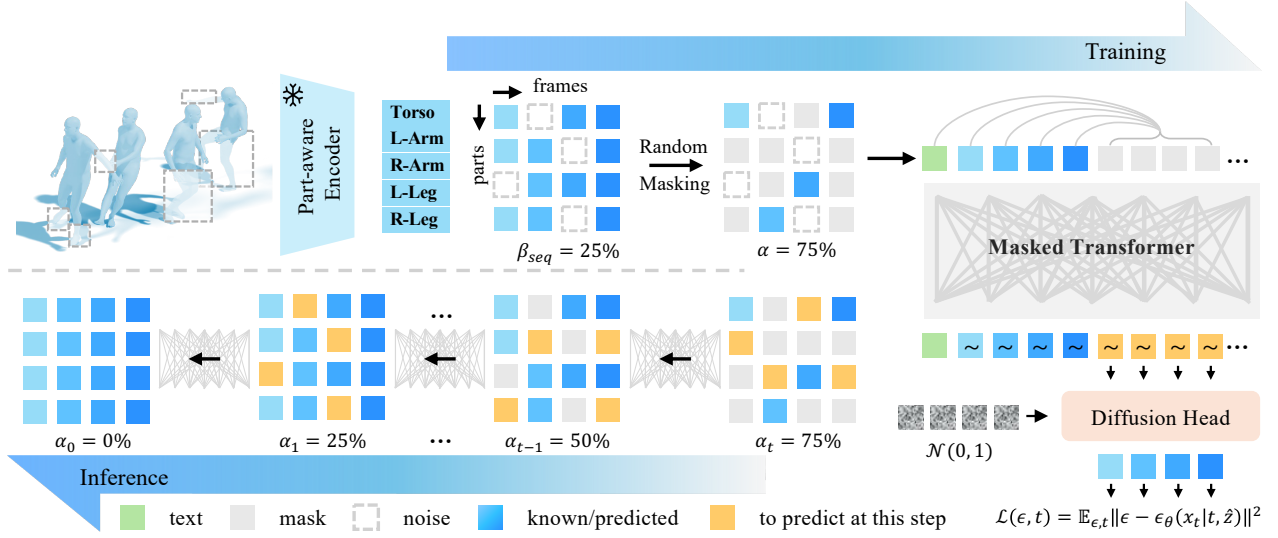


Figure 3: Overview of RoPAR. We first estimate which body parts are credible from 2D pose confidence, then encode only those parts into a part-aware latent grid. During training, non-credible tokens are always masked, a subset of credible tokens is additionally masked, and the generator predicts the missing tokens from text and visible context. A latent diffusion head refines the completed sequence before decoding.

The key design choice is that the reconstruction loss is applied only to credible parts:

$$\mathcal{L}_{\text{P-VAE}} = \sum_{i=1}^N \sum_{p \in P} O_i^p \left(\|m_i^p - \hat{m}_i^p\|_2^2 + \lambda_{\text{kl}} \text{KL}(q_\phi(z_i^p | m_i^p) \| \mathcal{N}(0, I)) \right). \quad (2)$$

When $O_i^p = 0$, the part contributes no reconstruction or KL gradient. As a result, the latent space is shaped by observed dynamics. HMR completions do not drive it. We share the encoder and decoder parameters across body parts, which keeps the latent representation comparable across limbs and reduces overfitting to any single part.

After encoding, each motion sequence becomes a latent grid $Z = \{z_i^p \mid i = 1, \dots, N, p \in P\}$ with one token per part per frame.

This tokenization has two practical effects. First, it localizes missing supervision: if only one arm is unreliable, the rest of the sequence is still usable. Second, it makes the later masking step semantically meaningful, because each missing token corresponds to a specific body part at a specific time. The masking pattern is no longer an arbitrary latent chunk.

3.3 Visibility-guided masked generation

This stage trains a text-conditioned generator to complete a part-by-time latent grid while avoiding supervision from unreliable parts. The generator receives the latent grid Z and the text prompt T . Non-credible tokens are always masked in the input because their recovered values should not be used as ordinary targets. To keep the learning difficulty comparable across sequences with different visibility ratios, we additionally mask a subset of credible tokens.

Let β_m be the fraction of non-credible tokens with $O_i^p = 0$ in a sequence, and let α be the desired total masking ratio. For credible

tokens, we sample a Bernoulli mask with probability

$$\gamma = \begin{cases} \frac{\alpha - \beta_m}{1 - \beta_m}, & \text{if } \beta_m < \alpha \text{ and } \beta_m < 1, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

When the non-credible fraction already exceeds the target masking ratio, we therefore set $\gamma = 0$ and do not additionally mask credible tokens. Sequences with no credible tokens contribute no term to $\mathcal{L}_{\text{gen}}$. The final input token $\hat{z}_i^p$ is

$$\hat{z}_i^p = \begin{cases} [M] & \text{if } O_i^p = 0, \\ [M] & \text{if } O_i^p = 1 \text{ and } r_i^p < \gamma, \\ z_i^p & \text{if } O_i^p = 1 \text{ and } r_i^p \geq \gamma, \end{cases} \quad (4)$$

where $r_i^p \sim U(0, 1)$ and $[M]$ is a learnable mask token.

The text-conditioned transformer G_θ predicts the masked credible tokens:

$$\mathcal{L}_{\text{gen}} = \sum_{i=1}^N \sum_{p \in P} 1[O_i^p = 1 \wedge \hat{z}_i^p = [M]] \|z_i^p - G_\theta(\hat{Z}, T)_{i,p}\|_1. \quad (5)$$

Tokens with $O_i^p = 0$ are therefore masked in the input but excluded from the training target. They create realistic missing-context patterns, and the model must infer them from text together with the remaining credible tokens. It is never trained to imitate weakly supported labels. In a large corpus, each body part appears as a credible target often enough that the model can learn these completions from text and cross-part context.

This corpus-level coverage matters because RoPAR does not require every clip to contain enough evidence to supervise a full motion. It is designed for datasets in which useful evidence is distributed unevenly across clips. A running action may expose the legs clearly in one clip and the upper body clearly in another. When that variation exists at scale, the model can aggregate these partial observations into a shared text-conditioned motion prior.

Table 1: Main K700-M results under two standard web-data policies. *Keep-All* retains the full recovered source pool and trains on the recovered full-body targets, while *Discard* removes any clip that contains a non-visible body part before training. Best results are in bold and second-best results are underlined. The Δ row reports RoPAR’s absolute improvement over the second-best method in each metric, with positive values always indicating better performance.

Methods	<i>Keep-All</i> (~200k)						<i>Discard</i> (~50k)					
	FID ↓	R1 ↑	R2 ↑	R3 ↑	Div. ↑	MM-D ↓	FID ↓	R1 ↑	R2 ↑	R3 ↑	Div. ↑	MM-D ↓
MDM [33]	37.812	0.487	0.651	0.770	6.421	17.986	35.274	0.481	0.644	0.764	6.083	18.214
MotionGPT [17]	30.954	0.525	0.683	0.793	6.884	17.259	31.406	0.519	0.676	0.786	6.512	17.541
MoMask [12]	24.334	0.585	<u>0.761</u>	0.834	7.551	15.827	<u>25.118</u>	0.582	0.754	0.828	<u>7.134</u>	15.968
MotionStreamer [36]	25.231	<u>0.591</u>	0.758	<u>0.836</u>	<u>7.902</u>	<u>15.459</u>	25.964	<u>0.599</u>	<u>0.765</u>	<u>0.839</u>	6.918	<u>15.881</u>
RoPAR (ours)	19.682	0.711	0.864	0.891	9.456	13.612	21.384	0.683	0.842	0.881	8.731	13.948
Δ	+4.652	+0.120	+0.103	+0.055	+1.554	+1.847	+3.734	+0.084	+0.077	+0.042	+1.597	+1.933

The transformer output is a coarse full-body latent grid. We refine it with a lightweight latent diffusion head:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x_0, t, \epsilon} \|\epsilon - \epsilon_{\theta}(x_t, t, G_{\theta}(\hat{Z}, T))\|_2^2, \quad (6)$$

where x_t is the noisy version of latent target x_0 at diffusion step t . This stage improves temporal smoothness and coordination at part boundaries.

The final training objective combines the three components:

$$\mathcal{L} = \mathcal{L}_{\text{P-VAE}} + \lambda_{\text{gen}} \mathcal{L}_{\text{gen}} + \lambda_{\text{diff}} \mathcal{L}_{\text{diff}}. \quad (7)$$

In practice, the three terms play different roles: $\mathcal{L}_{\text{P-VAE}}$ shapes the latent space, $\mathcal{L}_{\text{gen}}$ learns text-conditioned completion from partial observations, and $\mathcal{L}_{\text{diff}}$ sharpens the final latent trajectory.

Training. Training follows the same logic end to end. For each web clip, we first recover a 3D motion sequence with HMR and derive part-level observability masks from ViTPose confidence on the original frames. The part streams are then encoded by the shared P-VAE, but reconstruction and KL terms are applied only on credible parts. We next construct the masked latent grid by always masking non-credible tokens and randomly masking additional credible tokens. The transformer predicts only the masked credible targets from text and the remaining visible context, and the diffusion head refines the completed latent trajectory before parameter updates with the joint objective above. This training procedure preserves the structured partial observations in each clip while preventing unreliable parts from acting as direct supervision.

3.4 Inference

At test time, no visibility mask is available because the model receives only text. We therefore initialize a fully masked part-by-time latent grid and perform iterative masked generation. At each iteration, the transformer predicts the remaining masked locations, commits the most confident subset of part-by-time tokens, and feeds the updated grid back into the next round. Early iterations establish the coarse action pattern, while later iterations focus on lower-uncertainty local details. Once all latent locations are populated, the diffusion head refines temporal continuity and cross-part coordination, and the shared P-VAE decoder reconstructs the final full-body motion sequence.

This inference setup is consistent with training. During learning, the model repeatedly completes latent grids in which some part-by-time tokens are missing because the corresponding body parts are unobservable or intentionally masked. Inference is simply the limiting case in which all locations are initially missing. Visibility patterns vary across the corpus: legs are clearly observed in some clips, arms in others, and torso motion in many. Under that condition, the model can aggregate complementary partial evidence into a shared text-conditioned motion prior. Each individual clip no longer has to provide full-body supervision.

4 Experiments

4.1 Experimental setup

Datasets. We evaluate on both noisy web data K700-M, which is our web-scale dataset and evaluation testbed built from Kinetics-700 [6]. We process the videos with WHAM [30] and obtain 198,627 motion sequences. More than 70% of these sequences contain at least one body part that is not visible in the source frames. We use Gemini [31] to generate clip-level descriptions and apply the same post-processing pipeline to all clips. For K700-M, we fix a train/validation/test protocol, retrain the standard text-motion evaluator on the K700-M training split, and use the same captions, split, and evaluator for all compared methods. We therefore interpret K700-M primarily as a controlled in-domain comparison under a shared protocol.

Implementation details. For generation, we train all models with AdamW using learning rate 10^{-4} and $\beta = (0.9, 0.99)$. Training uses a 10k-step warm-up schedule, gradient clipping at 1.0, and 50k optimization steps. For evaluation on K700-M, we retrain one shared text-motion evaluator following TMR [28] on the training split using AdamW with learning rate 10^{-4} , batch size 256, and 100 epochs, and we use this same evaluator for every compared method. The evaluator uses a 128-dimensional motion latent and a 640-dimensional text latent, with 4 transformer layers, 4 attention heads, dropout 0.1, and feed-forward width 1024. Across all experiments, motion is decomposed into the same five body parts as in Sec. 3.1. The credibility threshold τ is fixed once by validation-time calibration on a held-out subset and then shared across all experiments.

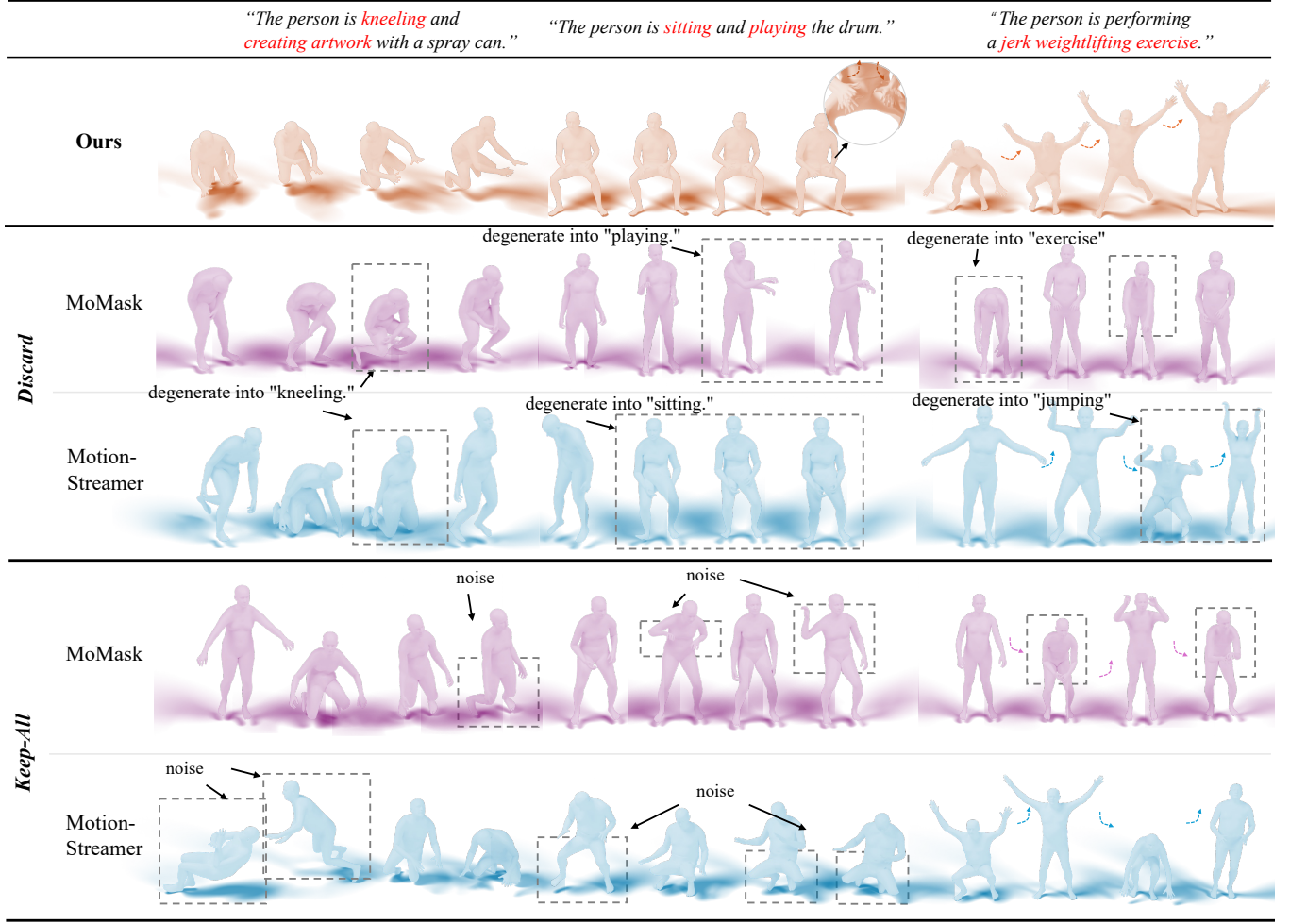


Figure 4: Qualitative comparison on K700-M under the *Keep-All* and *Discard* data policies. RoPAR is compared with MoMask and MotionStreamer on prompts where only part of the body is directly observed.

Metrics. We report the standard T2M metrics: Fréchet Inception Distance (FID), R-Precision (R1/R2/R3), Diversity, and multimodal distance (MM-D). Lower FID and MM-D are better, while higher R-Precision and Diversity are better.

4.2 Web-data policies

Our first experimental question is how a full-body T2M system should use web-recovered motion when some body parts are not directly observed. Existing T2M systems are built to align text with a single full-body motion target, whereas many K700-M clips contain recovered motions that are only partly credible. Before comparing architectures, we therefore have to decide how these recovered motions become supervision.

We study two standard web-data policies. In *Keep-All*, we retain the full recovered source pool and train on the HMR output as if the recovered full-body target were trustworthy everywhere. This policy preserves data scale and semantic coverage, but it also exposes the model to unsupported joints. In *Discard*, we remove

any clip that contains a non-visible body part and keep only the complete-visibility subset for training. This yields cleaner full-body targets, but reduces the effective training set to roughly one quarter of the original source pool and removes many hard web cases.

These are the only practical policies available to earlier full-body T2M systems on web data. If the model assumes a trustworthy full-body target, it must either trust the recovered full body or train only on clips that remain fully visible. This is also why part-aware models such as ParCo and FineMoGen are the closest prior architectures: only a part-decomposed representation can express supervision for one region of the body while withholding it from another. RoPAR uses that same structure differently. It uses parts to decide where supervision is allowed. We therefore report all K700-M comparisons under two matched settings: a smaller but cleaner setting produced by *Discard*, and a larger but noisier setting produced by *Keep-All*. Table 1 is an end-to-end comparison under those two web-data policies, while the two RoPAR entries also give a direct data-policy ablation of whether keeping the full web

source pool helps once unsupported joints are removed from direct supervision.

4.3 Comparison with baselines

We compare RoPAR with four strong T2M baselines. MDM [33] is a representative diffusion baseline, while MotionGPT [17], Mo-Mask [12], and MotionStreamer [36] represent language-style motion modeling, masked generation, and causal latent generation, respectively. FineMoGen [43] and ParCo [45] matter especially because they expose explicit body-part structure, but they are more informative in a focused partial-supervision comparison than in the generic baseline table. We therefore use Table 1 for the standard matched K700-M baselines, and later add a dedicated part-aware comparison in Table 2. In each K700-M setting of Table 1, the baselines are retrained on the same selected motion-text pairs as RoPAR and evaluated with the same protocol.

Table 1 reports results under the two matched K700-M settings introduced above: a smaller but cleaner *Discard* setting, and a larger but noisier *Keep-All* setting. Within each column, all methods use the same motion-text pairs, captions, split, and retrained K700-M evaluator. The differences come from training architecture and supervision policy. They do not come from a method-specific scoring backend. Because RoPAR is evaluated under both settings, the gap between its two entries is also a direct ablation of whether retaining the full web source pool helps our model.

The *Keep-All* column is where RoPAR matters most. It outperforms all strong baselines across every reported metric, reducing the best baseline FID from 24.33 to 19.68, improving the best baseline R@1 from 0.591 to 0.711, and increasing Diversity from 7.55–7.90 to 9.46. This is the web-data regime we care about most. Keeping the full source pool helps when the model can ignore unsupported joints. It fails when every reconstructed joint is treated as ground truth.

The *Discard* column makes a different point. Cleaner targets alone still leave a large gap. Removing incomplete clips gives the baselines better full-body supervision. It also shrinks the effective training pool to roughly one quarter of the original web source. Even in this cleaner regime, RoPAR remains strongest, reaching 21.38 FID, 0.683 in R@1, and 8.73 in Diversity. RoPAR is also worse under *Discard* than under *Keep-All*, which means that the removed clips still contain useful supervision once unsupported joints stop acting as targets. The practical takeaway is clear: keep the hard web clips and learn only from the parts that are actually supported by image evidence.

Only part-aware backbones can directly express clip-internal selective supervision, so we adapt ParCo and FineMoGen to the same credible-part masking rule and evaluate them in the scale-preserving *Keep-All* regime. We also include a RoPAR ablation without credible-only supervision to separate the benefit of selective supervision from the rest of the architecture. Table 2 shows that both part-aware baselines become much stronger once they are trained with partial supervision, which confirms that part decomposition is the right interface for this setting. Removing credible-only supervision from RoPAR also causes a clear drop across all reported metrics, while the full RoPAR model remains best overall. The result points to two ingredients at once: the part-aware architecture

Table 2: Part-aware comparisons on K700-M in the scale-preserving *Keep-All* regime. We adapt ParCo and FineMoGen to the same credible-part loss masking rule and include a RoPAR ablation without credible-only supervision to isolate the value of selective part-level supervision.

Methods	FID ↓	R1 ↑	R3 ↑	Div. ↑	MM-D ↓
ParCo w/ P-Sup.	23.914	0.641	0.854	8.742	14.926
FineMoGen w/ P-Sup.	21.876	0.669	0.874	8.981	14.287
RoPAR (ours)	19.682	0.711	0.891	9.456	13.612
RoPAR w/o P-Sup.	32.418	0.558	0.767	7.934	15.472

Table 3: Ablations on K700-M. The top block evaluates P-VAE reconstruction quality, and the bottom block evaluates the full generator.

Reconstruction	FID ↓	MPJPE ↓
P-VAE	0.21	6.95
w/o part-wise decomposition	1.86	19.83
w/o shared parameter	0.89	9.82
Generation	FID ↓	R1 ↑
RoPAR	19.68	0.71
w/o part-wise decomposition	21.36	0.58
w/o diffusion head	71.92	0.41

matters, and so does using credibility throughout latent learning, visibility-guided masking, and diffusion refinement.

As a fully observed control, we also evaluate RoPAR on HumanML3D, where all training parts are credible and the partial-supervision issue is absent. RoPAR remains competitive there and preserves strong retrieval-style alignment, which suggests that the architecture is not useful only in the target web-partial regime.

4.4 Ablations

Model components. This experiment tests whether each design choice is tied to the partial-supervision setting or is simply an isolated engineering trick. Table 3 validates both the P-VAE and the generator. Replacing the part-aware VAE with a full-body VAE increases reconstruction error substantially, which shows that part-level encoding is important even before generation. Removing shared parameters also degrades reconstruction quality, suggesting that a common latent geometry across parts is helpful.

The bottom half of Table 3 shows that the gains do not come from visibility filtering alone. Removing part-wise decomposition hurts both FID and R@1, because the model loses the explicit structure needed to separate credible and non-credible supervision. Removing the diffusion head causes a much larger drop, which indicates that latent refinement matters once the coarse sequence has been assembled from partial evidence.

Taken together, the ablations show that explicit part decomposition is necessary for partial supervision, and that the default five-part split provides a good trade-off between locality and coordination.

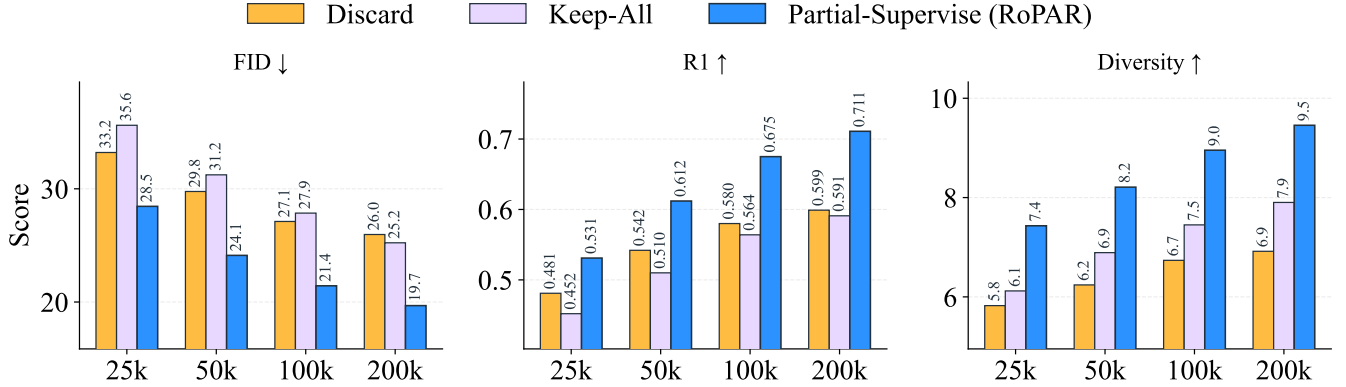


Figure 5: Scaling behavior under unreliable supervision across original recovered K700-M source pools of 25k, 50k, 100k, and 200k clips. The x-axis is indexed by the recovered source pool before policy-specific filtering; under *Discard*, the effective retained set is much smaller because clips with any non-visible body part are removed before training. RoPAR avoids the quality-coverage trade-off between *Keep-All* and *Discard* and achieves the best FID, R1, and Diversity at all four scales.

Scaling behavior. Figure 5 extends the comparison to recovered source pools of 25k, 50k, 100k, and 200k clips. All runs use the same source pools, captions, and evaluator, while *Discard* and *Keep-All* also share the MotionStreamer backbone and optimization settings. *Discard* provides cleaner targets and can improve FID and R1 at smaller scales, but filtering hard clips narrows coverage and consistently lowers Diversity. *Keep-All* preserves that coverage yet learns from unsupported joints, which hurts FID and R1. RoPAR avoids both failure modes and achieves the best FID, R1, and Diversity at every scale. The 200k endpoints agree with Table 1: larger web pools help only when unsupported joints are excluded from direct supervision.

4.5 Qualitative results

Figure 4 connects the numerical gains to case-level behavior. Mo-Mask usually captures the action category but is less stable in fine part coordination, especially for asymmetric arm use, seated balance, and grounded hand contact. MotionStreamer is more fluent, yet under *Keep-All* it inherits artifacts from unsupported recovered joints: the “spray can” motion shows local limb noise, “playing the drum” drifts toward an implausible seated pose, and weightlifting loses left-right consistency and temporal smoothness. These errors reflect the same supervision mismatch in different regions of the body: plausible HMR completions are treated as observed targets even when the source frames do not support them.

Under *Discard*, MotionStreamer sees cleaner targets but far fewer hard partial-visibility examples. Local noise can decrease, yet the reduced training coverage makes outputs more conservative and removes the structure needed for seated, crouched, or asymmetric actions. RoPAR avoids both failure modes by retaining these clips without treating unsupported joints as targets. This lets it learn from visible local dynamics across the full web pool while ignoring weak reconstructed labels, preserving asymmetric cues, contact geometry, and smooth part-to-part transitions more consistently across all three examples. The contrast mirrors the quantitative trade-off: *Keep-All* preserves action coverage but propagates unsupported

completions, whereas *Discard* suppresses those labels at the cost of rare poses and interactions. Selective part supervision retains the useful side of both policies.

5 Conclusion

We study text-to-motion learning from partially observed web motion. RoPAR keeps the full source pool but supervises only image-supported body parts through visibility-aware decomposition, part-aware latent modeling, and masked generation. On K700-M, this selective supervision improves motion quality and text alignment under both *Keep-All* and *Discard*. Its advantage across training scales further shows that partially useful clips are more valuable than either their unsupported joints or wholesale filtering, provided that credibility is modeled inside the objective rather than used only for preprocessing. The HumanML3D control additionally indicates that the architecture remains effective when all parts are credible.

More broadly, the bottleneck in web-scale T2M is not only recovery quality, but also the mismatch between what the video reveals and what the model is asked to imitate. Partial observability is therefore structured missing supervision rather than a binary data-cleaning problem. Aligning the training target with image evidence lets a model use difficult clips without learning prior-based completions as ground truth. The formulation also separates data utility from label completeness: a clip need not provide a trustworthy full body to contribute reliable local dynamics. This perspective may extend to other video-driven 3D generation tasks in which recovered signals combine direct observations with model-based inference.

Limitations. RoPAR depends on the calibration of its visibility estimator: false negatives discard useful supervision, while false positives retain weakly supported parts. Its fixed threshold may also require recalibration when the pose estimator or video domain changes. The current study targets single-person videos and does not explicitly address multi-person interaction or severe camera motion. Extending credibility estimation to these cases remains future work.

References

- [1] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. 2018. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE Int'l Conf. Robotics and Automation (ICRA)*. IEEE, 5915–5920.
- [2] Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2Pose: Natural language grounded pose forecasting. In *2019 Int'l Conf. 3D Vision (3DV)*. IEEE, 719–728.
- [3] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. 2022. TEACH: Temporal action composition for 3D humans. In *2022 Int'l Conf. 3D Vision (3DV)*. IEEE, 414–423.
- [4] Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, and Sonal Gupta. 2023. Make-An-Animation: Large-scale text-conditional 3D human motion generation. In *Proc. IEEE/CVF Int'l Conf. Computer Vision (ICCV)*. 15039–15048.
- [5] Bin Cao, Sipeng Zheng, Ye Wang, Lujie Xia, Qianshan Wei, Qin Jin, Jing Liu, and Zongqing Lu. 2025. A Real-Time Controllable Vision-Language-Motion Model. In *ICCV*.
- [6] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987* (2019).
- [7] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. 2023. Executing your commands via motion diffusion in latent space. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*. 18000–18010.
- [8] Filipe R. Cordeiro and Gustavo Carneiro. 2020. A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations?. In *2020 33rd SIGGRAPH Conf. Graphics, Patterns and Images*. IEEE, 9–16.
- [9] Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. 2025. Go to Zero: Towards zero-shot motion generation with million-scale data. *arXiv preprint arXiv:2507.07095* (2025).
- [10] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. 2021. Synthesis of compositional animations from textual descriptions. In *Proc. IEEE/CVF Int'l Conf. Computer Vision (ICCV)*. 1396–1406.
- [11] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proc. IEEE/CVF Int'l Conf. Computer Vision (ICCV)*. 14783–14794.
- [12] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. 2024. Momask: Generative masked modeling of 3d human motions. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*. 1900–1910.
- [13] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating diverse and natural 3d human motions from text. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*. 5152–5161.
- [14] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022. TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3D human motions and texts. In *European Conf. Computer Vision*. Springer, 580–597.
- [15] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proc. 28th ACM Int'l Conf. Multimedia*. 2021–2029.
- [16] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Adv. Neural Inf. Process. Syst.* 31 (2018), 8536–8546.
- [17] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. Motiongpt: Human motion as a foreign language. *Adv. Neural Inf. Process. Syst.* 36 (2023), 20067–20079.
- [18] Chuhao Jin, Haosen Li, Bingzi Zhang, Che Liu, Xiting Wang, Ruihua Song, Wenbing Huang, Ying Qin, Fuzheng Zhang, and Di Zhang. 2025. PlanMoGPT: Flow-enhanced progressive planning for text to motion synthesis. *arXiv preprint arXiv:2506.17912* (2025).
- [19] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. 2020. VIBE: Video inference for human body pose and shape estimation. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*. 5253–5263.
- [20] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J Black. 2021. PARE: Part attention regressor for 3D human body estimation. In *Proc. IEEE/CVF Int'l Conf. Computer Vision (ICCV)*. 11127–11137.
- [21] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proc. IEEE/CVF Int'l Conf. Computer Vision (ICCV)*. 2252–2261.
- [22] Junnan Li, Richard Socher, and Steven C. H. Hoi. 2020. DivideMix: Learning with noisy labels as semi-supervised learning. In *Int'l Conf. Learning Representations*.
- [23] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. 2023. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Adv. Neural Inf. Process. Syst.* 36 (2023), 25268–25280.
- [24] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *Proc. IEEE/CVF Int'l Conf. Computer Vision*. 5442–5451.
- [25] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. 2025. Rethinking Diffusion for Text-Driven Human Motion Generation: Redundant Representations, Evaluation, and Masked Autoregression. In *Proc. Computer Vision and Pattern Recognition Conf.* 27859–27871.
- [26] Mathis Petrovich, Michael J Black, and Gül Varol. 2021. Action-conditioned 3D human motion synthesis with transformer VAE. In *Proc. IEEE/CVF Int'l Conf. Computer Vision (ICCV)*. 10985–10995.
- [27] Mathis Petrovich, Michael J Black, and Gül Varol. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conf. Computer Vision*. Springer, 480–497.
- [28] Mathis Petrovich, Michael J Black, and Gül Varol. 2023. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proc. IEEE/CVF Int'l Conf. Computer Vision*. 9488–9497.
- [29] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. 2024. World-Grounded Human Motion Recovery via Gravity-View Coordinates. In *SIGGRAPH Asia Conf. Proc.*
- [30] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. 2024. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*. 2070–2080.
- [31] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05330* (2024).
- [32] Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. 2022. MotionCLIP: Exposing human motion generation to CLIP space. In *European Conf. Computer Vision*. Springer, 358–374.
- [33] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano. 2023. Human motion diffusion model. In *Eleventh Int'l Conf. Learning Representations*.
- [34] Ye Wang, Sipeng Zheng, Bin Cao, Qianshan Wei, Weishuai Zeng, Qin Jin, and Zongqing Lu. 2024. Scaling Large Motion Models with Million-Level Human Motions. *arXiv preprint arXiv:2410.03311* (2024).
- [35] Yuxin Wen, Qing Shuai, Di Kang, et al. 2025. HY-Motion 1.0: Scaling flow matching models for text-to-motion generation. *arXiv preprint arXiv:2512.23464* (2025).
- [36] Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. 2025. MotionStreamer: Streaming Motion Generation via Diffusion-based Autoregressive Model in Causal Latent Space. *arXiv preprint arXiv:2503.15451* (2025).
- [37] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. In *Adv. Neural Inf. Process. Syst.*
- [38] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. 2023. Decoupling human and camera motion from videos in the wild. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*. 21222–21232.
- [39] Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, Lei Yang, and Ziwei Liu. 2024. Whac: World-grounded humans and cameras. In *European Conf. Computer Vision*. Springer, 20–37.
- [40] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. 2023. Generating human motion from textual descriptions with discrete representations. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*. 14730–14740.
- [41] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2024. MotionDiffuse: Text-driven human motion generation with diffusion model. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 6 (2024), 4115–4128.
- [42] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. 2023. ReMoDiffuse: Retrieval-augmented motion diffusion model. In *Proc. IEEE/CVF Int'l Conf. Computer Vision (ICCV)*. 364–373.
- [43] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. 2023. FineMoGen: Fine-grained spatio-temporal motion generation and editing. *Adv. Neural Inf. Process. Syst.* 36 (2023).
- [44] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlec, Siyu Tang, and Federica Bogo. 2024. RoHM: Robust human motion reconstruction via diffusion. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*. 14606–14617.
- [45] Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. 2024. Parco: Part-coordinating text-to-motion synthesis. In *European Conf. Computer Vision*. Springer, 126–143.